Sport Trackers and Big Data: Studying user traces to identify opportunities and challenges

Rudyar Cortés, Xavier Bonnaire, Olivier Marin, Pierre Sens

Sport Trackers and Big Data: Studying user traces to identify opportunities and challenges

Rudyar Cortés, Xavier Bonnaire, Olivier Marin, Pierre Sens

RESEARCH REPORT
N° 8636
November 2014
Project-Teams ARMADA

ISSN 0249-6399 ISRN INRIA/RR--8636--FR+ENG

Sport Trackers and Big Data: Studying user traces to identify opportunities and challenges

Rudyar Cortés, Xavier Bonnaire, Olivier Marin, Pierre Sens

Project-Teams ARMADA

Research Report n° 8636 — November 2014 — 15 pages

Abstract: Personal location data is a rich source of big data. For instance, fitness-oriented sports tracker applications are increasingly popular and generate huge amounts of location data gathered from sensors such as GPS and accelerometers. Discovering new opportunities and challenges behind this kind of data requires knowledge about global user input in terms of volume, velocity, variety, and values. Gathering and analysing traces from a real world sports tracker service provides insight on these matters, but sport tracker services are very protective of such data due to privacy issues. We avoid this issue by gathering public data from a popular sports tracker server. In this paper, we present our database which is freely available online, and our analysis and conclusions from a big data perspective.

Key-words: Sport Trackers, Personal Location Data, Big Data.

RESEARCH CENTRE
PARIS – ROCQUENCOURT

Domaine de Voluceau, - Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex

French title

Résumé : Pas de résumé

Mots-clés : calcul formel, base de formules, protocole, différentiation automatique, génération de code, modélisation, lien symbolique/numérique, matrice structurée, résolution de systèmes polynomiaux

Contents

1 Introduction **3**

2 Opportunities with sport sensing data **4**
2.1 Sports-oriented social networks 4
2.2 Location Based Services 5
2.3 Global healthcare 5
2.4 Urban planning 5

3 Study Case: Endomondo Sports Tracker **6**

4 Evaluation **7**
4.1 Data Distribution 7
4.1.1 Distribution by year 7
4.1.2 Distribution by type of sport 8
4.1.3 Distribution by distance 8
4.1.4 Distribution by country 8
4.1.5 Distribution by continent 8
4.1.6 Distribution by duration 9
4.2 Estimation of a real world database 9

5 Discussion **10**
5.1 Big data characteristics 11
5.1.1 Data Volume 11
5.1.2 Velocity 11
5.1.3 Variety 12
5.1.4 Value 12
5.2 Big data challenges 13
5.2.1 Data cleaning 13
5.2.2 Data management 13
5.2.3 Scalable data mining and processing 13
5.2.4 Real time processing 13
5.2.5 Data privacy 13

6 Related Work **13**

7 Conclusion **14**

1 Introduction

The proliferation of smartphones [1] with location capabilities leads to the generation of massive amounts of personal location data, reaching the big data scale [2] [3]. The main source of this type of data is sensing devices such as GPS, that allow locating a smartphone within a geographical area.

From the big data perspective, applications that generate location data exhibit three main charasteristics: (i) the data originates from millions of users/sources; (ii) users generate a massive volume of small pieces of sensing data with spatio-temporal properties that requires continuous processing (iii) the data is mainly semi-structured. Such data can be seen as a massive flow of small pieces of spatio-temporal data generated by million of users, which must be captured, stored, indexed, and processed in several cases in real time [4].

In this context, health oriented sport sensing applications for smartphones have become extremely popular. Projects like *Endomondo* [5], *Nike+* [6], and *RunKeeper* [7] currently gather millions of users that generate billions of GPS data points as user traces. For example, the Strava Project [8] has stored and analyzed over 250 billion of GPS data points generated by sport activities. Recent projects like Fitbit [9] and Nike FuelBand [6] go further: they track all daily activities like sleep, diet habits, sports, and walking habits.

In order to identify the big data challenges behind this kind of applications, it is crucial to acquire real traces. These allow measuring and understanding the main big data characteristics associated with this type of data such as volume, velocity, and variety [10].

In this paper, we focus on acquiring and analyzing user traces from a popular sport tracker application.

The main contributions of this paper are the following:

- a study of our freely available dataset which covers information about 333,689 user traces from a real-world sport tracker application[11],
- and a discussion about the big data characteristics, opportunities and challenges behind this kind of application.

2 Opportunities with sport sensing data

The analysis of sensing data generated by sport trackers presents several opportunities in different fields such as sports-oriented social networks, location-based services, urban planning, and global healthcare [12, 13, 14, 15]. The present section provides some examples of the usage and analysis that this kind of data induces.

2.1 Sports-oriented social networks

The increasing popularity of sport tracker applications currently leads to the fast development of sports-oriented social networks. In a similar way to traditional social networks such as Facebook and Twitter, sports-oriented social networks allow users to share their sports-related performance and experience with friends and enrich their posts with comments and photos. A recent study [16] conducted over users of the MyFitnessPal sport tracker application shows that 54% of its users share their sport activities on Facebook and 56% prefer to exercise in a social context.

Examples of uses of sport data to create social interactions are the followings.

- **Social interaction based on geographical proximity**. Users that practise the same sport in a common geographical zone may wish to develop social interactions. For example, nearby users that were running at the same time might want to compare their routes and results.
- **Real time tracking**. Some applications like RunKeeper offer their premium users the possibility to track and share their route with friends in real time. Moreover, users can share pictures along the route in real time.
- **Social sport challenges**. Massive sport challenges consist in public goals related to a specific sport activity. Users that accept the challenge can be ranked according to their results. This aggregation of users around a common goal can induce new social interactions.

2.2 Location Based Services

The spatio-temporal properties of sport sensing data can be used to build location based services.

- **Searching for sport pals**. A sport pals search refers to the location of nearby places where there are people practising a defined type of sport. For example, a runner would like to discover nearby parks where there are others users running.

- **Route discovery**. Routes traversed by users represent a rich source of information for nearby users. For example, a user can be interested in discovering a new route that is 10 kms long and both starts and finishes near to the current position.

2.3 Global healthcare

The impact of big data in the healthcare industry is now a hot topic [2]. In this context, sport tracker data is very valuable and allows assessment of the physical activity of the population, and more generally of its well-being.

- **Global healthcare feedback**. One such field of application is *obesity control.* It is one of the main factors towards chronic diseases worldwide. Governments make significant efforts to keep their population healthy and out of hospital. For instance, the cost of the U.S. government's obesity prevention campaign is estimated between $147 billion to $210 billion per year [17]. The timely feedback of sport activities represent an efficient metric to measure the impact of government projects to prevent obesity in the population.

- **Postoperative monitoring**. Many surgical operations that compromise patient mobility require long recovery times. Hospitals monitor the physical activity and the daily movements of patients who underwent such surgery to determine their recovery status and to allow early diagnosis in case of complications. Currently this task is human resource intensive and subjective because it is performed as interviews. Sport tracker services can automate the monitoring and improve the objectivity of the assessment. They also represent a rich source of statistical data for hospitals. For example, a total knee replacement entails a recovery period between 7 and 12 weeks and the analysis of user mobility can help the surgeon acquire an accurate feedback with respect to the patient recovery status.

2.4 Urban planning

Comprehensive knowledge about the geographic distribution of sport activities and about user behaviors with respect to sports represents a useful source of information for urban planners and governments.

- **Management of sports facilities**. Large urban sports complexes usually regroup many facilities, and their management can be very costly. Studying the habits of the majority of people practising sports in the vicinity can provide very important insight on how to optimize expenses: by adapting opening and closing times, among other things.

- **Planning cycle routes**. Many cities have invested in a network of bike routes: asphalt trails separated from roads and traffic. Monitoring sports data produced by cyclists offers a cost-efficient feedback on the daily usage of each route, and provides a strong basis for future improvements of the network.

Parameter	**Value**
Type of sport	Running
Start date and time	2011-09-11 14:39:00
Distance	14.16 km
Duration	1h24m:19s
Max speed	4.42 [min/km]
Average speed	6.01 [min/km]
Calories	1179

Table 1: Example of workout resume

Parameter	**Value**
Longitud	−79.975199921716
Latitude	32.777506706718
Altitude	−14.42
Distance	0.12614343 [km]
Duration	51081 [ms]
Pace	8.44095 [min/km]

Table 2: Example of GPS tuple generated every Δt.

3 Study Case: Endomondo Sports Tracker

This section describes our study case: the Endomondo sports tracker application. Endomondo is a popular social sports tracker application with more than 20 millions of users worldwide in 2013 [18]. It allows users to track and share their workout results with friends, and to publish them for everyone to see in the case of public profiles. Thus, public workouts are a rich source of real world traces regarding sports activities.

The mobile application uses the GPS sensor to track the device along its route and it registers incremental parameters like distance, speed, duration and time. When the workout is finished, it can be enriched with user comments and shared with the Endomondo social network or with traditional social networks like Facebook and Twitter.

Gathering traces from sports activities

In order to acquire real world traces for our study, we gathered all public information from the Endomondo web server. The data set covers the whole sports activities history from $15,090$ users chosen randomly; it covers a total of $333,689$ workouts obtained over five months in 2014.

The first step was to code a script to download a raw log file containing all workouts registered by every user. Every entry in the log file contains the *user profile*, *workout summary*, and *GPS trace*. The user profile contains general information regarding the user, such as *username*, *country*, *birthdate*, *postal code*, *sex*, *weight*, and *height*. The workout summary presents statistics assessed during the session, such as *maximum speed*, *average speed*, theoretical amount of burnt *calories*, and *weather*. Table 1 shows an example of a workout summary.

The GPS trace contains all GPS points generated along the route of the workout session every Δt. The value of Δt is typically some seconds and it depends on the type of sport. Table 2 shows an example of a GPS tuple generated along a route.

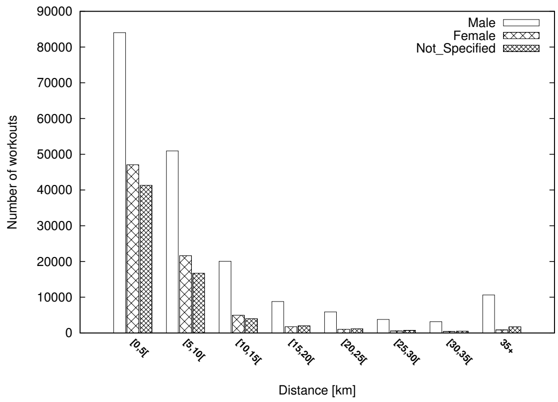

Figure 1: Data distribution in terms of distance traversed in workouts of 15,090 users

Users	$15,090$
Workouts	$333,689$
Male	$6,269$
Female	$2,812$
Not specified	$6,009$
Types of sports	69
GPS rows	$56,851,893$
Database Size	$5.71[GB]$

Table 3: Sport tracker database summary

Once we extracted all raw logs, we dumped them into a PostgreSQL database consisting of three tables: *user*, *workout*, and *GPS trace*.

Table 3 shows a summary of our sports activity database. $6,009$ users (39.82%) did not provide information about their gender. Note that the tracking of a workout route generates a significant volume of data in terms of GPS tuples. A workout session averages 170 GPS tuples, which represents a size of 18.13 [KB].

4 Evaluation

This section presents our analysis of public workout traces extracted from the Endomondo sports tracker server. Our goal is to evaluate the data distribution, the data volume, and data generation velocity produced by this type of applications.

4.1 Data Distribution

4.1.1 Distribution by year

In order to evaluate the activity of our sample comprising 15,090 users, we measured the number of active users among them per year between 2007 and 2010 and the number of workouts that they published. We define an active user as a user that publishes at least one workout in the

Year	Users	Workouts
2007	2	99
2008	3	186
2009	10	206
2010	483	3,729
2011	2,361	25,535
2012	5,952	88,711
2013	**9,706**	**161,337**

Table 4: Number of workouts registered per year by 15,090 users

Sport	Workouts	Percentage [%]
Running	**142,265**	**42.63**
Cycling	76,329	22.87
Walking	66,585	19.95
Others	48,510	14.53

Table 5: Distribution by type of sport

year. We omit 2014 because the gathered data set would not span the full year. Table 4 shows the detailed results. The main observation we make is that the number of users and the number of published workouts increases strongly every year. This fact is consistent with the observation of the increasing popularity of this kind of applications.

4.1.2 Distribution by type of sport

Table 5 shows the number of registered workouts per type of sport. Endomondo's most popular sport activity is running, which represents 42.6% of the workout publications, followed by cycling and walking.

4.1.3 Distribution by distance

In order to describe the workout data distribution by distance, we measured the number of sport activities generated in increasing ranges of 5 kilometres. Figure 1 shows the distribution of workout distances. Most users publish workouts that remain under 5 kilometres.

4.1.4 Distribution by country

Table 6 shows the user distribution by country. The country with the highest number of registered users is the U.S., which accounts for 18.5% of all users, followed by Spain and the United Kingdom.

4.1.5 Distribution by continent

Table 7 shows the user distribution by continent. The continent with the highest number of users is Europe, which accounts for 60.84% of all users, followed by America and Asia.

Country	Users	Percentage
United States	**2802**	**18.5**
Spain	2092	13.8
United Kingdom	1271	8.42
Denmark	1242	8.23
Poland	1166	7.72
France	531	3.51
Others +98	5988	39.66

Table 6: Users distribution by country

Continent	Users	Percentage
Europe	**9181**	**60.84**
America	3665	24.28
Asia	1740	11.53
Others	504	3.33

Table 7: Distribution of users by continent

Users	Size [GB]	Workouts	GPS tuples
15 x 10^6	7.5	$439,363$	$74,691,710$
30 x 10^6	15.1	$878,727$	$149,383,590$
50 x 10^6	25.3	$1,464,546$	$248,972,820$

Table 8: Estimation of the average size of the data received in a single day

4.1.6 Distribution by duration

Figure 2 shows the distribution of workouts by duration. Our main observation is that most workouts last less than 1 hour.

In order to validate the quality of our sample data we compared our results with public information released by Endomondo in June 2012 [19] when they had a database composed of 20 million users. We were pleased to find that our sample data matches the released results. For example, they show that the most popular activity was running (44%) followed by cycling (22%) and walking (17%). These results are very similar to our results measured from our sample data.

4.2 Estimation of a real world database

In this section we estimate the current volume and velocity of sensing data from sport activities. In 2014, there are roughly 30 million Endomondo users[18]. In order to conduct our evaluation we measure the flow of data from our sample population and we scale it to the magnitude of the real user base. As sports activities registered by Endomondo keep increasing every year, we use the activity published in 2013 to conduct our estimations because we feel it is the best representation of the current sports activity.

Figure 3 presents the estimated number of workouts generated per month by 30 million users. The maximum number of workouts was reached in august with $37,558,648$ workouts and the minimum in february with $16,510,934$ workouts. From our sample data we estimate that a single

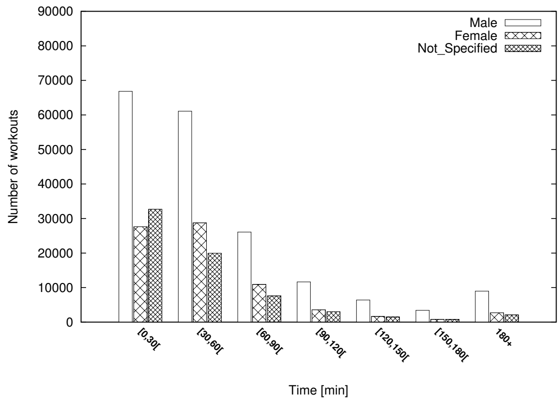

Figure 2: Workout distribution by duration

workout generates an average of 170 GPS tuples. Following this result, the estimated number of GPS tuples generated in a single month is between 2.8 and 6.3 billions.

Figure 4 shows the estimated number of workouts generated per day by 30 million of users. We estimate that the peak of workout generation in a single day is close to 1.4×10^6 workouts and it is reached several times in periods of high activity. On the other hand, during periods of low activity, we estimate that the number of workouts generated is around $400,000$ workouts per day.

In order to estimate the generation frequency of data from sport activities in a single day, we analysed the activity generated in days with high and low activity in 2013. As our sample data provides no timezone information, we used the Google Timezone API [20] to associate every user to a defined timezone. For every user, we took a GPS point from a single workout and we associated it to a defined timezone. Then, we translated all workouts to the UTC timezone. Finally, we used the start time and the incremental time parameters inside every GPS tuple to calculate the number of tuples generated per second in a single day.

Figure 5 shows an estimation of the maximum insertion frequency generated in a single day. The maximum number of tuples generated at the same time was around $25,000$ tuples/second. Figure 6 shows the minimum insertion frequency registered in a single day. In this case the maximum frequency was around $10,000$ tuples/second.

Table 8 shows an estimation of the average number of workouts received in a single day with respect to the number of users. Every day, a single sport application with 30 million users receive an average data size of 15.1 [GB] composed of $878,727$ workouts, for a total of $74,691,710$ GPS tuples.

5 Discussion

In this section we discuss the main big data characteristics and challenges pertaining to sports tracking applications.

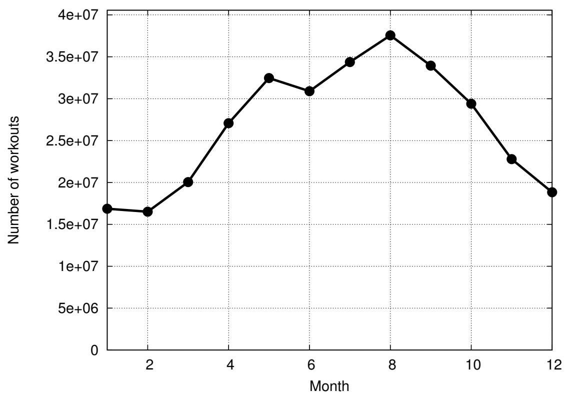

Figure 3: Estimation of the number of workouts generated per month by 30 million users

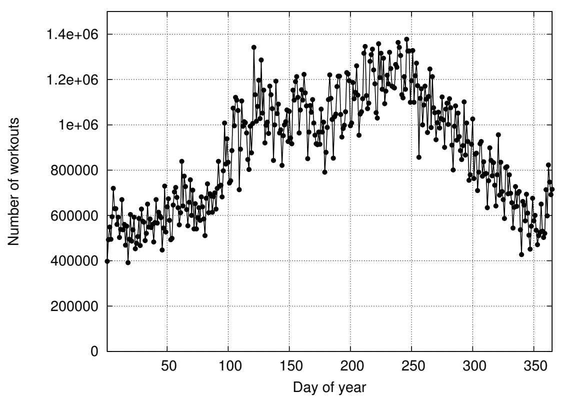

Figure 4: Estimation of the number workouts generated per day by 30 million users

5.1 Big data characteristics

5.1.1 Data Volume

In terms of storage size, the raw GPS data generated in a single day is in the gigabytes order. This is relatively small compared to applications that store photos or videos because the size of a single GPS tuple is in the bytes order, whereas photos or videos are in the megabytes order. However, sport tracker applications generate a high data volume in terms of the number of objects (GPS tuples) that must be stored and processed.

5.1.2 Velocity

The velocity of the generation of data from sport trackers is variable and reaches its maximum value when most people practises sports at the same time. From our evaluation we estimate that an application with 30 million users reaches a peak of roughly $25,000$ tuples/second. Moreover, the recent adoption of new technologies that tracks all daily activities, and the use of social networks to share sport activities suggests that the velocity of the generation of location data

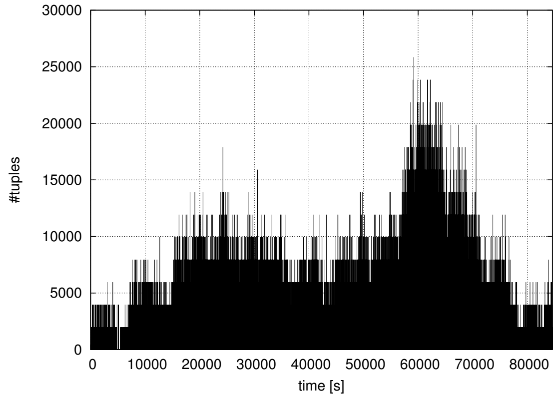

Figure 5: Maximum GPS data generation frequency registered in a single day by 30 million of users

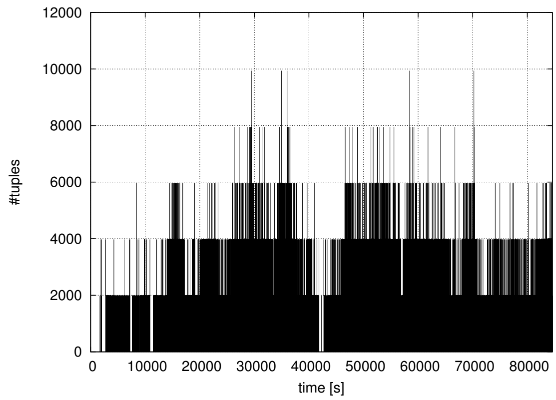

Figure 6: Minimum GPS data generation frequency registered in a single day by 30 million of users

will continue increasing.

5.1.3 Variety

Location data generated from smartphones is mainly structured data. However, the current trend of sports-oriented social networks enriches location data with comments and photos related to sports activities. This new way of interacting around sports gives a new unstructured dimension to sport sensing data.

5.1.4 Value

The spatio-temporal properties of personal location data and its relatively small size compared to other applications suggest that this kind of data offers a higher value per byte stored than other applications. It is possible to extract value of this kind of applications in different fields such as global healthcare, urban planning, sport social networks, and location based services.

5.2 Big data challenges

5.2.1 Data cleaning

Sensor data from GPS sensors can be noisy due to precision or hardware problems. Moreover, *data incompleteness* due to the limited battery life of smartphones can be observed. Examples of this kind of problems include routes do not tracked or sport activities tracked outside of roads. A key challenge here is to design efficient methods to detect and filter this kind of data.

5.2.2 Data management

The massive volume of sensor data presents several issues in terms of data management. The main challenge here is how to efficiently organise this kind of data to extract value. The spatio-temporal properties of data must be considered. For example, how to index data to be able to search closest routes performed by others in a tolerable time?

5.2.3 Scalable data mining and processing

The high data volume in terms of the amount of GPS tuples generated from sport activities imposes scalability challenges to data mining algorithms. It could be interesting here to explore *in-network processing*, where some computations can be performed before to submit the data to the storage level.

5.2.4 Real time processing

Real time processing is a key challenge of big data. The always increasing number of users that are adopting sport trackers to practise sports present several scalability challenges to centralised servers. For example, *nearby activity detection* requires to track all users that are practising sports near to a given position. This kind of queries can be performed by millions of users at the same time overwhelming the processing capacity centralised servers. A key challenge here is to propose distributed architectures to manage location data in real time.

5.2.5 Data privacy

Data generated from sport trackers contains sensitive personal location data (e.g. The starting point of a route could reveal the user address). A key challenge here is to propose techniques that protect the user privacy when the data is analyzed.

6 Related Work

The increasing popularity of sport trackers has recently attracted the research community to understanding and analyzing this kind of data. An initial discussion about the secondary use of sensor data and the emergence of a conceptual architecture is provided in [13]. The main focus of current research in sports-oriented data is to understand its geographical distribution. A recent work [14] identifies popular sports areas in individual cities by using sensor data gathered from the Nokia Sport Tracker service. Another study [21] extracts and analyses sensor data from the MapMyFitness sport tracker server in order to describe the use of sports related facilities in Winston-Salem, USA. A full discussion about how this kind of data can be used to acquire an effective feedback on physical activity and how it can be used for global healthcare is provided in [15]. The evaluation of different architectures for continuous processing of GPS data is provided [22] . Some studies [14] [21] provide general statistics regarding sports activity data sets. However, none of these studies focus on assessing the big data characteristics pertaining to such applications like data volume, velocity, and variety of sensor data generated by sports trackers.

7 Conclusion

The analysis of user traces from sport trackers allows us to explore new opportunities and challenges behind this kind of sensor data. In this paper we present our dataset composed of user traces from sports activities worldwide, which is freely available online. We provide an estimation of the current volume and velocity of the generation of sensor data pertaining to this type of application, and we discuss the main opportunities and challenges from a big data perspective. We show that the current data volume generated by these applications can easily reach billions of tuples, and we describe the dynamic data flow generated in a single day. As a consequence of this study, we are working on a distributed architecture to meet the main big data challenges presented in this paper.

References

[1] P. E. Ross, "Top 11 technologies of the decade," *IEEE Spectr.*, vol. 48, no. 1, pp. 27–63, jan 2011. [Online]. Available: http://dx.doi.org/10.1109/MSPEC.2011.5676379

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[3] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen *et al.*, "The mobile data challenge: Big data for mobile computing research," in *Pervasive Computing*, no. EPFL-CONF-192489, 2012.

[4] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, pp. 1–39, 2014.

[5] "Endomondo sports tracking [online]," Available: http://www.endomondo.com, accessed: 2014-04-07.

[6] "Nike+ sports tracking [online]," Available: http://nikeplus.nike.com, accessed: 2014-04-07.

[7] "Runkeeper sports tracking [online]," Available: http://runkeeper.com, accessed: 2014-04-07.

[8] "Strava sports tracking [online]," Available: http://labs.strava.com, accessed: 2014-04-07.

[9] "Fitbit project [online]," Available: http://www.fitbit.com, accessed: 2014-04-07.

[10] L. Douglas, "3d data management: Controlling data volume, velocity and variety," *Gartner. Retrieved*, vol. 6, 2001.

[11] "Sport tracker dataset [online]," Available: https://cloud.lip6.fr/public.php?service=files&t=22b1f338d0da8b92a87adb5b13703624.

[12] C. C. Aggarwal and T. Abdelzaher, "Integrating sensors and social networks," in *Social Network Data Analytics*. Springer, 2011, pp. 379–412.

[13] A. Clarke and R. Steele, "How personal fitness data can be re-used by smart cities," in *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on*. IEEE, 2011, pp. 395–400.

[14] L. Ferrari and M. Mamei, "Identifying and understanding urban sport areas using nokia sports tracker," *Pervasive and Mobile Computing*, vol. 9, no. 5, pp. 616–628, 2013.

[15] L. Atallah, B. Lo, and G.-Z. Yang, "Can pervasive sensing address current challenges in global healthcare?" *Journal of epidemiology and global health*, vol. 2, no. 1, pp. 1–13, 2012.

[16] "Myfitnesspal report [online]," Available: https://blog.myfitnesspal.com/2014/05/myfitnesspal-ebook-the-rise-of-the-fitness-tribe/, accessed: 2014-04-07.

[17] J. Cawley and C. Meyerhoefer, "The medical care costs of obesity: an instrumental variables approach," *Journal of health economics*, vol. 31, no. 1, pp. 219–230, 2012.

[18] "Endomondo blog [online]," Available: http://blog.endomondo.com/2013/10/16/endomondo-fitness-app-runs-past-20-million-users-and-reaches-profitability/, accessed: 2014-04-07.

[19] "Endomondo infographics [online]," Available: http://blog.endomondo.com/tag/infographic/, accessed: 2014-04-07.

[20] "Google timezone api [online]," Available: https://developers.google.com/maps/documentation/timezone/, accessed: 2014-04-07.

[21] J. A. Hirsch, P. James, J. R. Robinson, K. M. Eastman, K. D. Conley, K. R. Evenson, and F. Laden, "Using mapmyfitness to place physical activity into neighborhood context," *Frontiers in public health*, vol. 2, 2014.

[22] S. Duarte, D. Navalho, H. Ferreira, and N. Preguiça, "Scalable data processing for community sensing applications," *Mobile Networks and Applications*, vol. 18, no. 3, pp. 357–372, 2013.

RESEARCH CENTRE
PARIS – ROCQUENCOURT

Domaine de Voluceau, - Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399